Originally published in: European Journal of Applied Linguistics vol. 7 (2019), No. 2, pp. 305-331

Harald Lüngen and Laura Herzberg*

Types and annotation of reply relations in computer-mediated communication

https://doi.org/10.1515/eujal-2019-0006

Abstract: This paper presents types and annotation layers of reply relations in computer- mediated communication (CMC). Reply relations hold between post units in CMC interactions and describe references from one given post to a previous post. We classify three types of reply relations in CMC interactions: first, technical replies, i.e. the possibility to reply directly to a previous post by clicking a 'reply' button; second, indentations, e.g. in wiki talk pages in which users insert their contributions in the existing talk page by indenting them and third, interpretative reply relations, i.e. the reply action is not realised formally but signalled by other structural or linguistics means such as address markers '@', greetings, citations and/or Q-A structures. We take a look at existing practices in the description and representation of such relations in corpora and examples of chat, Wikipedia talk pages, Twitter and blogs. We then provide an annotation proposal that combines the different levels of description and representation of reply relations and which adheres to the schemas and practices for encoding CMC corpus documents within the TEI framework as defined by the TEI CMC SIG. It constitutes a prerequisite for correctly identifying higher levels of interactional relations such as dialogue acts or discussion trees.

Keywords: reply relations, corpus annotation, computer-mediated communication, CMC, Text Encoding Initiative, TEI

Zusammenfassung: Der vorliegende Artikel stellt Typen und Annotationsebenen von Antwortrelationen in der internetbasierten Kommunikation (IBK) vor. Antwortrelationen bestehen zwischen Posts in IBK-Interaktionen und beschreiben Referenzen, die zwischen einem Initialbeitrag und einem Folgebeitrag bestehen. Wir klassifizieren drei Arten von Antwortrelationen in IBK-Interaktionen: erstens, technische Antwortrelationen, welche dadurch gekennzeichnet sind, dass durch das Betätigen einer „Antwort"-Schaltfläche eine Antwort initiiert wird, bspw. in Blogs; zweitens,

***Corresponding author: Laura Herzberg,** Universität Mannheim, Germanistische Linguistik, Schloss, D-68131 Mannheim, Germany, E-Mail: herzberg@uni-mannheim.de
Dr. Harald Lüngen, Leibniz-Institut für deutsche Sprache, R5 6-13, D-68161 Mannheim, Germany, E-Mail: luengen@ids-mannheim.de

Publikationsserver des Leibniz-Instituts für Deutsche Sprache
URN: http://nbn-resolving.de/urn:nbn:de:bsz:mh39-92645

Einrückungen, z. B. auf Wikipedia-Diskussionsseiten, in denen Benutzer ihre Beiträge in die entsprechende Stelle des Diskussionsverlaufs einfügen, indem sie ihre Beiträge einrücken und drittens, interpretative Antwortrelationen, bei denen die Antwort nicht formal realisiert wird, sondern durch andere strukturelle oder linguistische Mittel signalisiert werden, wie z. B. dem Adressierungsmarker „@“, Begrüßungs- und Verabschiedungsformeln, Zitaten und/oder Frage-Antwort-Strukturen. Wir analysieren die bestehenden Praktiken bei der Beschreibung und Darstellung solcher Relationen in Korpora und geben Beispiele für Chat, Wikipedia-Diskussionsseiten, Twitter und Blogs. Anschließend präsentieren wir einen Annotationsvorschlag, der die verschiedenen Ebenen der Beschreibung und Darstellung von Antwortrelationen kombiniert und sich an die Praktiken zur Kodierung von IBK-Korpusdokumenten innerhalb der Text Encoding Initiative (TEI), wie sie von der TEI CMC SIG definiert wurde, hält. Die Annotation von Antwortrelationen stellt eine Voraussetzung für die korrekte Identifizierung höherer interaktionaler Ebenen, wie z. B. die Klassifizierung von Dialogakten oder Baumstrukturen, dar.

Stichworte: Antwortrelationen, Antwortstrukturen, Korpusannotation, internetbasierte Kommunikation, IBK, Text Encoding Initiative, TEI

Resumen: Este documento introduce tipos y capas de anotación de las relaciones de respuesta en la comunicación mediada por ordenador (CMC). Las relaciones de respuesta se mantienen entre las unidades de mensaje de las interacciones de CMC y describen referencias de un mensaje dado a un mensaje anterior. Clasificamos tres tipos de relaciones de respuesta en las interacciones de CMC: primero, las respuestas técnicas, es decir, la posibilidad de responder directamente a un mensaje anterior usando el botón “responder”; segundo, hendiduras, por ejemplo, en las páginas de discusión de Wikipedia en las que los usuarios insertan sus contribuciones en la página de conversación existente al indentarlos, y la tercera, relaciones interpretativas de respuesta, es decir, la acción de respuesta no se realiza formalmente, sino que se señala por otros medios estructurales o lingüísticos, como los marcadores de dirección ‘@’, saludos, citas y/o estructuras de pregunta y respuesta. Vamos a mirar a las prácticas existentes en la descripción y representación de tales relaciones en los corpus y ejemplos de chat, páginas de discusión de Wikipedia, Twitter y blogs. A continuación, proporcionamos una propuesta de anotación que combina los diferentes niveles de descripción y representación de las relaciones de respuesta y que se adhiere a los esquemas y prácticas para codificar documentos de corpus CMC dentro del marco TEI, tal como se define en el TEI CMC SIG. Esto forma un prerrequisito para identificar correctamente los niveles más elevados de relaciones interaccionales, como los actos de diálogo o los árboles de discusión.

Palabras clave: relaciones de respuesta, anotación de corpus, comunicación mediada por computadora, CMC, Text Encoding Initiative, TEI

1 Introduction and motivation

In this paper, we examine the nature of various types of "reply", "addressing", or "reference" relations that exist between post units in computer-mediated communication (CMC) and which describe a reference from one given post to a previous post.

Reply relations appear when users respond to each other in their posts. Since interactions in CMC are versatile, depending on the genre and topic, reply relations do not only hold in question-answer structures. Reply relations include any kind of reactions that occur when two users interact with each other, e.g. when user A and B do not share each other's opinions and ideas and they disagree, when user B simply comments on a post by user A or when user B gives an explanation to a topic without user A asking directly, etc. We think that annotating reply relations constitutes a prerequisite for correctly identifying relational structures between user contributions and it is also a step towards improved annotations for higher levels of interaction analysis such as dialogue acts (Ferschke et al. 2012) or discussion trees (Laniado et al. 2011). We classify three types of reply relations in CMC interactions: *technical replies*, *indentations*, and *interpretative reply relations*. Our goal is to sort out the different levels of description and annotation that are involved, and to propose a solution for their combined representation within the TEI annotation framework. We adhere to the TEI Special Interest Group (SIG) on CMC, in which solutions for representing CMC corpus documents have been developed either by defining good practices for using elements from the regular TEI, or by customising CMC-specific new elements and attributes (Beißwenger et al. 2016).

By example of the studies by Holmer (2008), Laniado et al. (2011) and Ferschke et al. (2012), we would like to show and motivate why an analysis and annotation of reply relations can improve the analysis of social interaction in written CMC.

Holmer (2008) investigated chat transcripts, aiming at displaying interaction structures in chats. Figure 1 shows an extract of a chat he gives as an example. He manually inserted references indicating which messages refer to which other message, e.g. the message with the order identification 3 from the user *Pink*, "yep.", is an answer that refers back to the question from user *Black* with the order ID 1.

After manually referencing the messages, Holmer (2008) derived communication threads from the structure of the relating messages in order to display the social interaction. The identification of references between chat messages offers an important key to the analysis of chat communication. Annotating these references is useful for the overall understanding of the relations between chat messages.

Reference ID	Order ID	Message
1	1	Black: Did you see that new Mel Gibson movie - I think it is called "Payback"?
2	2	Pink: I saw the academy awards last night. Did you watch it?
1	3	Pink: yep.
3	4	Pink: It was very violent, but funny.
3	5	Black: You saw it? You liked it?
2	6	Black: How did it end up - who won?
1	7	Red: I heard it was good.
6	8	Pink: It was OK. At least Titanic didn't win everything.

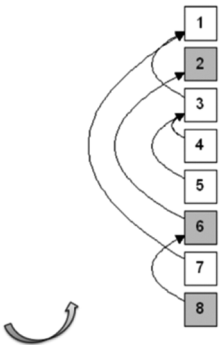


Figure 1: Display of reply relations in chat (from Holmer 2008: 4, 6).[1]

Such reply relations have been described in the literature for other CMC genres besides chat.

Name in Spanish [edit]

Why is the Spanish name of Antwerp here??? I understand the French name because a part of Belgium speak French.

> Belgium was part of the Spanish Netherlands before, that's why. --Michael (talk) 12:44, 19 September 2008 (UTC)
>
> > Come on, that's silly. It hasn't been ruled by Spain in 300 years. Spanish is no longer a relevant language in this city. 12.345.67.890 (talk) 00:25, 11 April 2010 (UTC)
> >
> > > It's not all that silly. The Spanish language may not be relevant in Antwerp (nor is French, which is spoken in a different part of Belgium), but Spanish rule plays a crucial role in the history of the city. (Antwerp remained Spanish during the Dutch revolt, see the articles introduction.) Mo (talk) 15:49, 19 January 2012 (UTC)
> > >
> > > > To give an example of why it's not silly, the man who designed the cathedral at Granada in Andalucia (Spain), trained with the St. Lucas guild in Antwerp. There was an enormous exchange of art and culture between the Southern Netherlands and Spain during the reigns of Charles V and Philip II. Any serious student of art history would need to know that Amberes means Antwerp. Felix (talk) 23:13, 13 July 2012 (UTC)

Figure 2: Display of indented posts in the Wiki talk page *Antwerp*.[2]

1 All presented examples that are not directly cited from the indicated literature have been pseudonymized by the authors.

2 https://en.wikipedia.org/wiki/Talk:Antwerp (accessed 12 January 2019).

Figure 2 shows a thread of the Wikipedia talk page of the article *Antwerp*. Its overall topic is stated in the thread heading which is then followed by an initial post and more contributions. The users reply to previous contributions by indenting their posts. This yields a tree structure, i.e. a hierarchy of talk page contributions, which can be exploited to detect structural patterns of interaction.

Laniado et al. (2011) model the interactions on a Wikpedia talk page as "discussion trees", where the root note corresponds to the talk page as a whole and the child nodes represent either user posts or structural elements such as sub-pages, headings, or sub-headings. The links between the user posts (called "comments" by Laniado et al. 2011) are built on the post indentations (interpreted as reply relations). A top-level post is linked to its heading, sub-headings are linked to their headings one level up, and the top-level headings are linked to the root node (representing the page as a whole). In Figure 3, their discussion tree which was automatically derived from the discussion page *Presidency of Barack Obama* is presented.

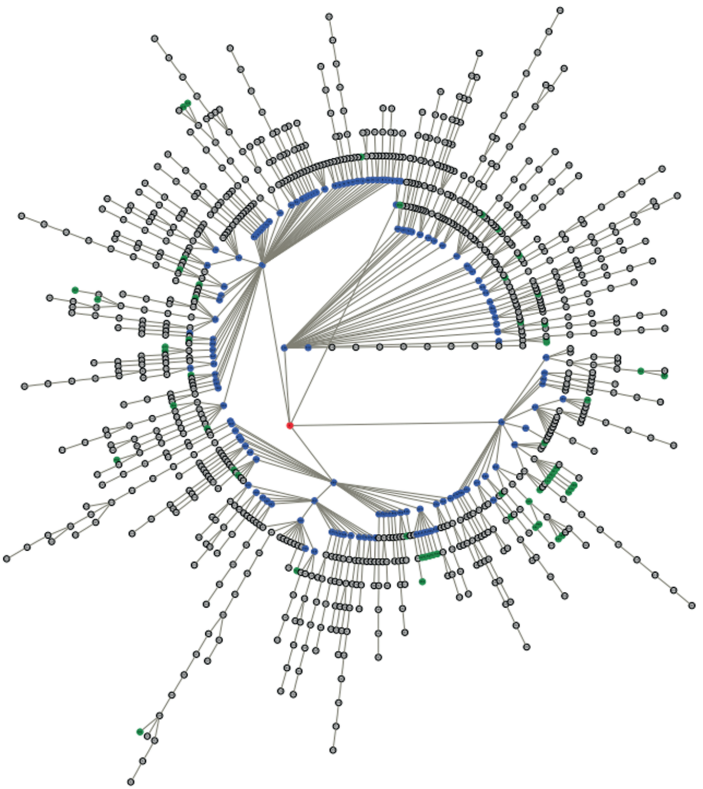


Figure 3: Display of a discussion tree from the discussion page *Presidency of Barack Obama* (source: Laniado et al. 2011: 180).

The red node represents the root, i.e. the article, blue nodes are structural elements, green nodes are anonymous posts (not signed with a proper user signature but only with an IP address) and grey nodes are posts from registered users. By employing such modes of visualisation, the different shapes of discussions are made explicit. There are posts that are placed directly after a structural node (e.g. a headline) and do not receive any replies; and large chain-like subthreads of posts, containing a sequence of replies between several users. Many chainlike sub-threads indicate the number of controversies in an article discussion. Laniado et al. (2011) also define a measure of the depth of a Wikipedia discussion based on these chains. Again, patterns of interaction are captured – this time by using the means of indentation. However, the indentation level does not always correspond to the interpretation of addressing cues, i.e. the concept of indentation cannot be taken for granted in all wiki discussion threads.

In addition to discussion trees, Wiki talk data have also been analysed in the framework of dialogue act classification. The basis of dialogue act classifications is the idea that utterances do have a certain function within a dialogue, especially in relation to the adjacent utterances. Since posts in CMC communication show features of spoken language, e.g. elliptical phrases, abbreviations or colloquial expressions, classifying dialogue acts can also be applied to posts.

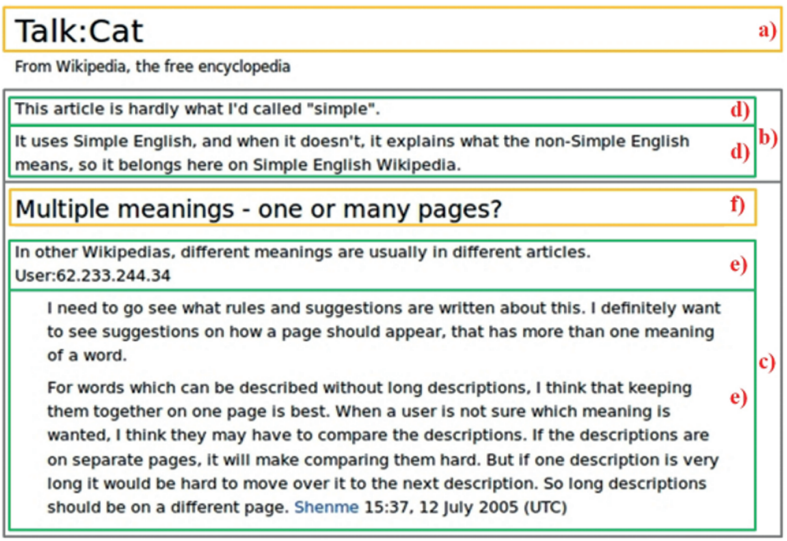


Figure 4: Display of a structured talk page according to a dialogue act classification: a) Talk page title, b) untitled discussion topic, c) titled discussion topic, d) unsigned turns, e) signed turns, f) topic title (source of example and annotation: Ferschke et al. 2012: 779).

Ferschke et al. (2012) aimed at analysing the content of Wiki talk pages with regard to investigating how the users coordinate their work in order to improve the articles. They developed 17 tags for Wiki talk pages by using 100 talk pages of the Wikipedia Simple English language version as a data basis. Each heading or post, presented in Figure 4, is labelled according to its specific function within the thread, e.g. in terms of its information content, labels such as "Information providing", "Information seeking" or "Information correcting" are applied. Nevertheless, dialogue act classification neglects the relational aspects between posts. Using reply relations between posts in such an annotation scheme would represent information that is highly relevant for the dialogue acts.

To summarize, we think that higher-level analyses such as the analysis of social interaction, the derivation of discussion trees, and the annotation of dialog acts could be improved by first reconstructing the "correct" reply relations and making them explicit.

The rest of the paper is structured as follows: The following section gives an overview of the types of reply relations that we identified on account of the way they are signalled in CMC as well as a survey of the existing practices in the description and annotation of such relations in Twitter, chat, Wiki talk, and blog corpora. In Section 3, we describe default and overriding effects of reply relations. Section 4 presents our proposal for a CMC annotation scheme within the TEI framework, which provides strategies for annotating the reply relations. In Section 5, we resume our main findings and discuss perspectives for future work.

2 Types of reply relations in written CMC

2.1 Technical reply

The most obvious and unambiguous type of reply relation can be observed in CMC genres where the client software, which is used to send a message (post) to the CMC platform, offers the possibility to reply directly to a previous post by clicking a button that is associated with the post and labelled 'reply' (or similar) (Figure 5, "Antworten" – 'reply').[3]

3 Examples are presented in their original language. For understanding purposes, an English translation is added in italics.

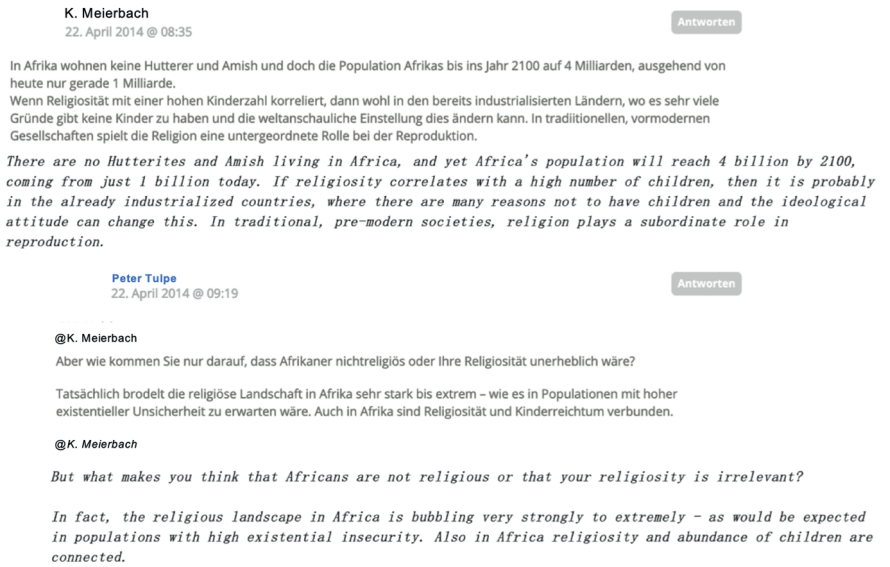


Figure 5: Display of a technical reply relation in blog comments of the blog *Nature of Belief*.[4]

It can be activated to start the process of composing and eventually sending the reply, and it represents the standard reply action available in CMC genres such as email, Usenet news, YouTube, or blog comments. Generally, the reply relation (which message replies to which) will also be documented in the metadata of the message, for example the "References" field in the NNTP header (Schröck and Lüngen 2015). We label this type of reply relation a *technical reply*. Technical reply relations frequently form reply chains, and, since several replies can be directed to the same previous message, the characteristic thread structures of such interactions arise. A post that is sent to the server without invoking a technical reply simply starts a new thread. CMC clients frequently display threads as indented list structures based on the reply ("References") information in the message protocol, (e.g. in the email client Thunderbird or in web browsers) via an HTML representation using nested lists or divisions (Figure 6).

4 https://scilogs.spektrum.de/natur-des-glaubens/die-anthropodizee-frage-wer-himmel/ (accessed 05 January 2019).

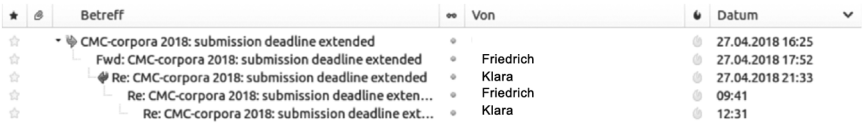


Figure 6: Display of technical reply relations in an email thread.

Figure 7: Display of technical relations in tweets.[5]

Figure 7 displays reply relations between different tweets and introduces Twitter as another CMC genre in which replies are initiated technically. In Figure 7, a representation of the initial tweet by the user *@Klara* and the first two replies to it are illustrated. The initial tweet is shown from the perspective of Klara's Twitter timeline. The presented view becomes visible once one clicks on the original tweet and the tweet itself, as well as all responses, load in a new window. Besides the options of liking and retweeting a tweet, users can directly reply to a tweet. In

5 https://twitter.com/[Klara]/status/1039850104663015424 (accessed 23 January 2019).

its default setting, the technical reply allows for addressing all mentioned users, i.e. those referenced with @ in the original tweet. The address @ is part of every username in Twitter. The first response by the user *@sarah_b_w* is a reply to *@Klara*. It continues the topic of emoji use by broadening its scope on images in graphics interchange format (gif) and asking about scientific research in the field. The second tweet by the user *@wahrwiss_GE* addresses this topic by providing links to papers that deal with this topic. The user *@wahrwiss_GE* sticks to the default reply setting as her tweet replies to all of the users mentioned in the two previous tweets from *@sarah_b_w*, *@Klara* and *@XBK_2018*.

2.2 Indentation

A second type of reply relations is represented by the indentation structures found on Wiki talk pages. Talk (or discussion) pages serve as a platform where Wiki authors coordinate their work and share ideas about edits and improvements to the associated Wiki article. From a technical point of view, talk pages are ordinary Wiki pages, just like the articles. Traditional Wiki software does not offer message or comment posting using a technical reply action as sketched under Section 2.1. Instead, users are instructed to insert their contributions in the existing talk page and to indent and sign them properly using the Wiki markup language[6] (Figure 8).

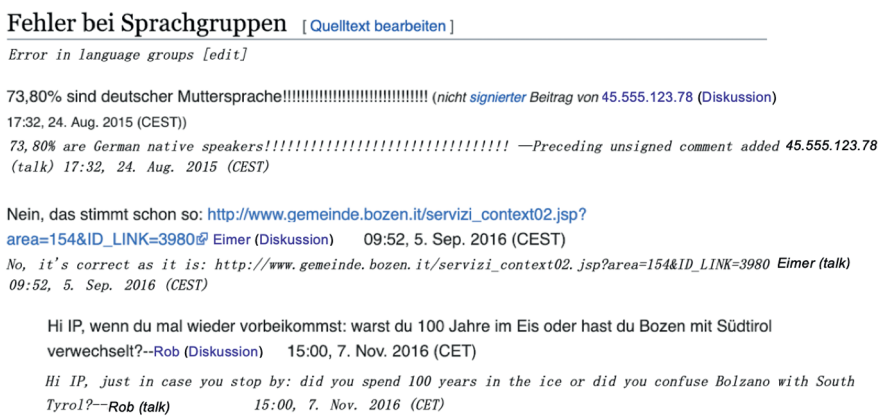


Figure 8: Display of the indentation on the Wiki talk page *Bolzano*.[7]

6 https://en.wikipedia.org/wiki/Help:Talk_pages#Indentation (accessed 12 January 2019).
7 https://de.wikipedia.org/wiki/Diskussion:Bozen (accessed 05 January 2019).

The sending action then always involves the sending of the whole, updated Wiki page to the server. Clearly, the indentation policy serves to imitate a threaded reply structure as known from the layout of CMC with technical reply, and as a result, the collaborative dialogues look like discussion threads on the web page. With respect to reply relations, a talk contribution (likewise called a *post* in the CMC corpora literature) is by default interpreted to indicate a reply to the post that is one level higher in the indentation hierarchy (Laniado et al. 2011, Margaretha and Lüngen 2014, Poudat et al. 2014, Ho-Dac et al. 2016).

2.3 Interpretative reply relations

Besides technical replies and indentations, we observe that relations between posts in CMC which researchers have identified as referencing or replying (e.g. Holmer 2008), can also be signalled by other structural or linguistics means. A good example of CMC where such alternative signalling abounds is chat. We regard chat as the sum of communication events that are realized using a specific chat technology which enables synchronous (non-simultaneous) exchange between several users based on a distribution method using the Internet Protocol (IP) and the client-server principle of the Internet as infrastructure (definition by Beißwenger 2007). This includes Internet relay chats, web chats as well as newer forms of chat systems, such as WhatsApp, Facebook-Messenger, and messenger systems that are implemented in social networking services, e.g. in Skype, Instagram, or Twitter. Neither in the composition of chat messages nor in the display of a chat log are technical replies or indentation structures applied. Hence, in chat, other indicators of the users' replying or addressing intentions are used, such as:

- a user name in combination with the address marker @ as in *@James* (default reading: this post is a reply to the most recent post by James);
- a name in combination with a greeting (*Hi Henry*, *Hello Linda*);
- simply a name (*Linda*);
- citation: explicitly quoting a piece of a previous post (most common in forum or email communication, often supported by the client software) (Schröck and Lüngen 2015; Grumt Suárez et al. 2016);
- Q-A structures: giving an answer to a question raised in a previous post.

Following Holmer's (2008) notation, who manually tagged references between chat messages and then used an application called *ChatLine* in order to visualise message and interaction structures (Holmer 2008), we also added a reference identification column to our Example 1:

Example 1: Excerpt of the chat *degu-chat_18-03-2003*.

Reference	Order #	User	Message
	27	Mausi	test farbe gewechselt *test colour changed*
	28	Maja	henry: ihr franken seid ja eh so ein völkchen *grins* nicht bös gemein, mein freund ist ja auch einer *henry: you Franconinans are a race apart *smile*mean no harm, my boyfriend is one as well*
	29	Lena	@Maja *gg* *@Maja *gg**
27	30	henry	so schön bunt hier *nicely colourful here*
28	31	henry	looooo l @Maja mein mann ist niederbayer es ist immer wieder zu schön *looooo l @Maja my husband is Lower Bavarian it's too good over and over again*
30	32	Camile	ich versuch es mal in rot *I'm trying red*
31	33	Lena	das wird ja richtig multikulturell... badner, franken, bayern.. *that will be really multicultural... people from Baden, Franconia, Bavaria..*
...			
	36	Maja	und mein lieblingsfrankenwort ist die gombadibilidäd *and my favorite franconian word is the gombadibilidäd*
	37	Mausi	die schwaben sind aber auch so ein völkchen für sich *but the swabians are also a race apart*
33	38	Chiara	Lena: Und eine Ex-Österreicherin, bitte! *Lena: And an ex-Austrian, please!*
...			
	40	Lolez	da gibts spätzle bei den schwaben *There's spaetzle in Swabia.*
...			
37	42	Lena	Mausi nur damit das gleich klar gestellt ist... ich bin badner *draufbesteh* *Mausi just so it's clear... i am badner *insist on it**

The column *Reference* displays the number of the referred post, i.e. post (30), as we can see it in Example 1, refers to post (27). The column *Order #* describes the position of the post within the thread. This example of the Dortmund chat corpus shows an unmoderated free time chat with several users. The excerpt shows that

the users are dealing with two topics simultaneously. First, they discuss the changing of their font colours and second, as they plan to meet, they exchange their experiences about South German regions and their inhabitants. Again, the chat's inherent reply structure becomes obvious: By the time user *henry* reacts (31) to *Maja*'s post (28) other posts have emerged and the users have to deal with them too. In comparison to blogs/forums and Wiki discussions, chat servers do not present formal structures of indentation or reply functions. The messages are displayed according to the order of their arrival at the chat server. In this genre, the users solely rely on linguistic cues, such as direct addressing with @ of a certain user as we have seen in Example 1.

There are more indicators, but these tend to get more implicit and ambiguous, e.g., when taking a closer look at Example 1, we can infer, because of the topic continuation, that (30) refers to (27) as well as (32) to (30). Also, a use of the pronoun *Du* (*you*) (not in the example) would signal a direct address to another user and consequently to one of her previous posts, but based on its form alone it cannot be decided who the addressee actually is. Humans no doubt often infer reply relations by understanding and interpreting that the content of a message forms a response or reaction in some sense to the content of a specific previous message, even without overt indicators.

Example 2 shows a chat extract from a WhatsApp group chat of the *Mobile Communication Database 2* (*MoCoDa 2*).[8] The participants talk about recommendations for a hairdresser.

Example 2: Excerpt of the WhatsApp group chat *9Gz4s*.[9]

Reference	Order #	User	Message
	1	Maren	Kann einer von euch vll einen . Guten Friseur empfehlen :) ? *Can any of you recommend a good hairdresser?*
1	2	Anne	Oliver Schmidt ist aber etwas teurer *Oliver Schmidt but it's a little more expensive*
1	3	Christina	Ich hab leider auch keinen . Gehe immer nur zu einer Freundin zum schneiden aktuell *I don't have one either. Currently, always going to a girl friend to get a cut*

8 *MoCoDa 2* is an ongoing project at University of Duisburg-Essen. The goal of this project is to create a database with a web frontend for repeated, donation-based collection of CMC messages, such as WhatsApp. For further information, see https://db.mocoda2.de/#/c/home (accessed 24 January 2019).

9 https://db.mocoda2.de/#/view/9Gz4s (accessed 24 January 2019).

Again, in the composition of WhatsApp messages, no technical replies or indentation structures are applied. As we can observe in Example 2, the indicators are more ambiguous as no user is directly addressed. Nevertheless, there are interpreted reply relations identifiable: The initial post by the user *Maren*, a question, is answered by two other chat participants – *Anne* and *Christina*. So both answers refer back to the initial question, forming Q-A structures.

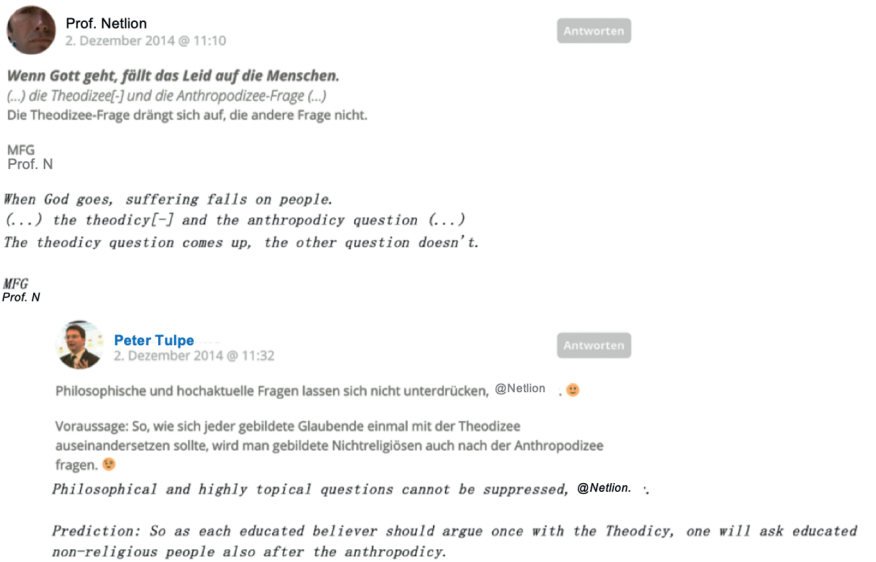
3 Default and overriding reply relations

Obviously, linguistic reply markers are also used in CMC genres that already offer a formal reply strategy (technical reply or indentation). Looking at examples from different CMC genres, we can interpret the formally signalled reply relation as the one that is valid by default, while it might or might not be "overridden" by a linguistically marked reply relation. We distinguish the following cases for posts that include a formal reply marker and at the same time a linguistic reply marker:

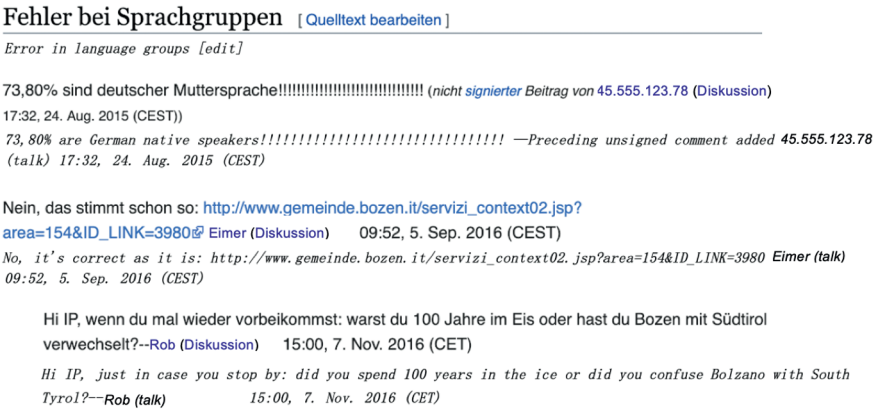
- Case A: the linguistic reply marker indicates the same reply relation that is already marked by the formal reply and hence re-inforces it;
- Case B: the reply relation indicated by the linguistic marker overrides the reply relation marked by the formal reply;
- Case C: the linguistic marker introduces an additional reply relation besides the one signalled by the formal reply.

In Example 3, the user *Tulpe* has used the reply button to answer to *Netlion*'s message. At the same time, he uses an address marker to address *Netlion*. Both strategies establish the same reply relation, i.e. the technical reply is reinforced.

In the German Wiki talk page on the city of Bolzano (Example 4), we have a discussion of the section on language groups in the articles, and we can identify three posts, i.e. user contributions. By the indentation, it seems that post 3 is a reply to post 2. But by the greeting and usage of the word *IP*, it becomes clear that post 3 is actually directed at post 1 (the unsigned message marked with an IP address). So the formal reply strategy in this example, i.e. the indentation, has apparently been applied erroneously and should be dropped in the interpreted reply structure. Especially in Wiki talk, this case frequently occurs. Even though there are generally accepted conventions of how to reply to previous postings, not all users do follow them and, for example, stick to the given level of indentation (Laniado et al. 2011). In Wiki discussions, the indentation level does not always correspond to the interpretations of addressing cues within a post, as illustrated in Example 4. According to Laniado et al. (2011), an unindented post can be interpreted as another initial post which relates to the overall topic that is stated in the



Example 3: Case A – the linguistic marker indicates the same reply relation as the formal reply.[10]



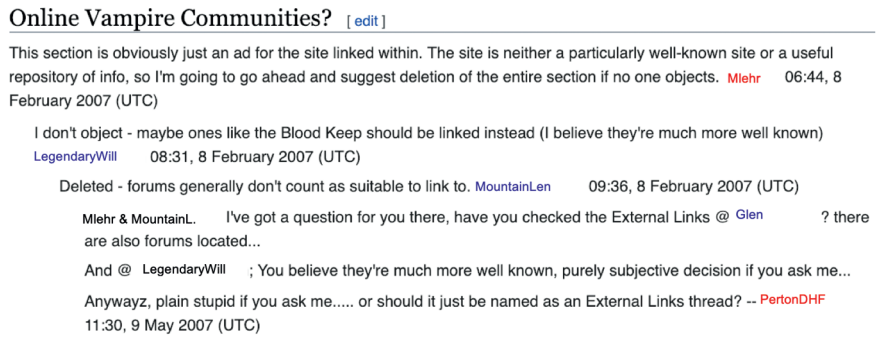
Example 4: Case B – the reply relation indicated by the linguistic marker overrides the reply relation marked by the formal reply.[11]

10 https://scilogs.spektrum.de/natur-des-glaubens/wenn-gott-leid-menschen-von/ (accessed 05 January 2019).

11 https://de.wikipedia.org/wiki/Diskussion:Bozen (accessed 05 January 2019).

thread heading. Analysing linguistic cues of reference show that this concept cannot be taken for granted in Wiki discussion threads. To verify this impression, we quantitatively analysed 300 posts of three different Wiki talk pages[12], i.e. 100 posts per talk page, in order to examine how many of the indentations that are inserted by the authors correspond to the interpretative cues in the posts, i.e. in how many of the posts does the level of indentation coincided with the content. Since Wikipedia authors usually have their own style in writing, we decided to analyse posts from different areas, such as politics, science and leisure, to minimise the possibility of indentation preferences that are specific to one user. The results confirm our assumption: the indentation level does not always correspond to the interpretation of addressing cues. In 32 % (i.e. 96 posts) of the 300 investigated posts, the displayed indentation level does not match the interpretative reply relation, e.g. even though some users were directly referring to each other, they did not indent their posts accordingly. From looking at the indentation level, posts that are not indented are interpreted as initial posts which relate to the overall topic of the thread, whereas in reality, the respective posts do refer to another one. Additionally, users may have their own writing styles leading to different practices: two users who were participating actively in the discussion on the article "The Legend of Zelda" rarely indented their posts – this eventually led to a ratio of 53 % falsely to 47 % correctly indented posts within the 100 investigated posts. In comparison, in the discussion "Flüchtlingskrise in Europa ab 2015" (*Refugee crisis in Europe since 2015*), a greater number of users was involved, leading to a ratio of 8 % falsely to 92 % correctly indented posts. This quantification indicates that the indentation level does not always correspond to the interpretations of addressing cues within a post.

12 We analyzed the first 100 posts of the following Wikipedia talk pages: "Flüchtlingskrise in Europa ab 2015" https://de.wikipedia.org/wiki/Diskussion:Fl%C3%BCchtlingskrise_in_Europa_ab_2015/Archiv/2, "The Legend of Zelda", https://de.wikipedia.org/wiki/Diskussion:The_Legend_of_Zelda/Archiv/1 and "Psychoanalyse", https://de.wikipedia.org/wiki/Diskussion:Psychoanalyse/Archiv/004 (all accessed 05 June 2019).



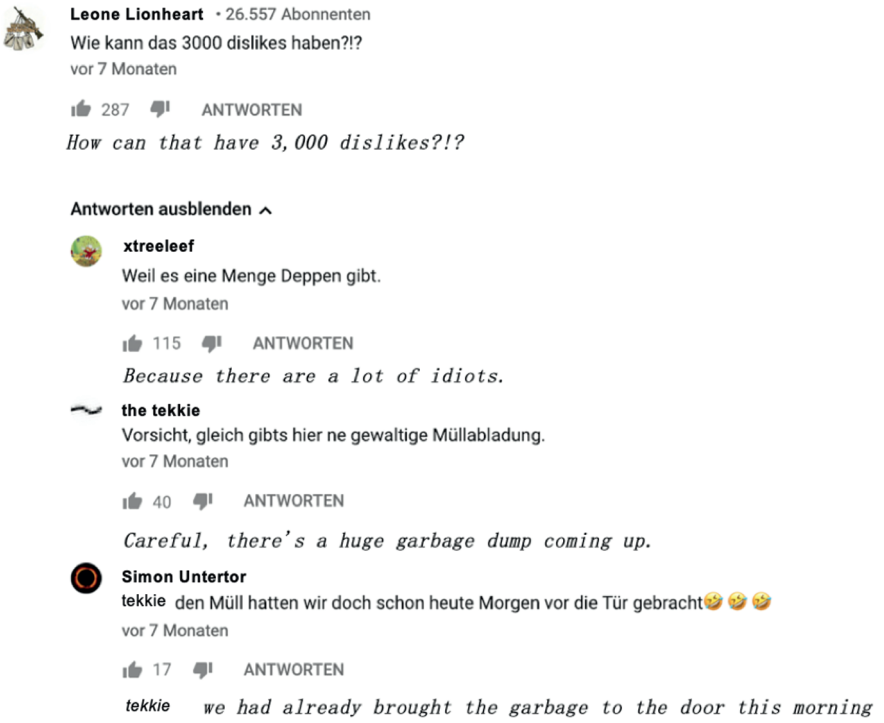
Example 5: Case C – the linguistic marker introduces an additional reply relation besides the one signalled by the formal reply.[13]

Example 5 displays another Wikipedia example, from the discussion of the article on *Vampire Counts*. By the formal indication, again the indentation, post 4 seems to be a reply to post 3. But by the names and the addressing terms used in post 4, we understand that post 4 is actually directed at each of the three previous posts at the same time. If we want to be even more precise, the first paragraph of post 4 is a reply to post 1 and post 3 (composed by the users *Mlehr* and *MountainLen*), and the second paragraph of post 4 is a reply to to post 2 composed by user *LegendaryWill*. If we keep analysing the post as a whole, unsegmentable unit, post 4 as a whole functions as a reply to each of the three previous posts.

Besides wiki talk, we observe that on many platforms that offer technical reply, there is a still a limitation of displayed indentation levels, which seems to regularly cause users to resort to linguistic markers of address to signal the reply status of a message.

In Example 6 from YouTube, only one level of indentation for the comments is displayed. Address markers, names, and other linguistic cues indicate reply relations between individual comments. In Example 6, the fourth post is not a reply to the first post as the displayed indentation might suggest, but to the third post.

13 https://en.wikipedia.org/wiki/Talk:Vampire_Counts_(Warhammer) (accessed 05 January 2019).



Example 6: Limitation of displayed indentation levels in YouTube comments.[14]

Blog platforms can also set limitations. Even though it is possible that an initial comment underneath a blog post can trigger a large follow-up discussion, their might be a limit on displayed indentation levels (Grumt Suárez et al. 2016). *Peter Tulpe*'s blog post "Die Anthropodizee-Frage. Wer den Himmel leerräumt, schafft die Menschheit ab[15]" ('The Anthropodize question. Whoever liberates heaven abolishes humanity') that is fully described in Grumt Suárez et al. (2016)[16], is an example of a changing blog platform. In 2014, when the author wrote this post and other users began to comment on it, the blog had an indentation levelling of

14 https://www.youtube.com/watch?v=gyal9T_fQ-8&t=50s (accessed 05 January 2019).

15 https://scilogs.spektrum.de/natur-des-glaubens/die-anthropodizee-frage-wer-himmel/ (accessed 05 January 2019).

16 The presentation from Grumt Suárez et al. (2016) that presents the indentation levels in the framework of the previous blog software is available via http://nl.ijs.si/janes/wpcontent/uploads/2016/09/CMC4NLP_Ljubljana_20092016_Grumt-Suarez_Karlova-Bourbonus_S.pptx (accessed 05 January 2019).

up to five. If we take a look at the current representation of the website, we can see that the maximum indentation display level is now two – so even though the underlying data from 2014 still includes the original, technical reply relations of the blog comments, the software limits now the display to a maximum of two indentation levels on the webpage.

4 Annotation proposal

We propose that a CMC annotation scheme provides different annotation strategies for annotating a.) the technical reply references as sketched under Section 2.1 above and documented in the protocols of email, Usenet, or blog comments, b.) the indentation structure as represented in Wiki text markup or HTML as known from Wiki talk (Section 2.2), and c.) the more interpretative reply structures induced by linguistic markers as sketched under Section 2.3. We generally propose a separate annotation layer to represent the interpretative, final reply structure at a more abstract level, which would combine reply relations of all three kinds where all cases of overriding are resolved.

4.1 TEI – Text Encoding Initiative

Our proposal adheres to the TEI framework, and within this framework, specifically to the proposals by the Special Interest Group on CMC (TEI CMC SIG), where solutions for representing CMC corpus documents using the TEI have been developed.

The TEI (Text Encoding Initiative) provides models and guidelines for the encoding of texts in the humanities such as manuscripts, critical editions, lexica, and also speech and language corpora. The TEI is run by an international consortium and publishes the so-called TEI Guidelines (current version: P5) which contain formal declarations of more than 500 XML elements and attributes, together with prose descriptions of their semantics (TEI Consortium, 2019). One can say that the TEI Guidelines define a de facto standard for text encoding in the humanities. In the guidelines, elements and attributes (such as <p> for paragraph, or <u> for utterance) are thematically grouped in modules called e.g. “header”, “drama”, “verse”, or “corpus”. These can be addressed when building subsets and validatable schemas (customisations) from the TEI. Since TEI encoding is based on the XML standard, it is fully software-independent. It is also highly community-driven, mostly through its lively mailing list, the journal jTEI, the annual TEI conferences, and through the various Special Interest Groups (SIGs) such as the TEI SIG for linguists, or the TEI Correspondence SIG.

Listing 1: TEI speech corpus encoding, using <u> and , for *utterance* and *division* (from Schmidt 2011).

```
<div>
    <u who="#SPK0">
        <anchor synch="#T6"/>Ah oui?. <anchor synch="#T7"/>
    </u>
</div>
```

Listing 2: Text corpus extract using , <head> <p>, and <s>, for *division*, *heading*, *paragraph*, and *sentence* of the Wiki article page "Alfred Hitchcock".

```
<div n="4" type="section" >
    <head><s> Paramount </s></head>
    <p><s>Die Erfahrung mit dem aufgezwungenen 3D-Verfahren zeigte Hitchcock die Grenzen
        bei Warner Brothers.</s>
    [...]
```

In the present version of the Guidelines, TEI P5 (TEI Consortium 2019), the bulk of which has first been published in 2007, no features for encoding the peculiarities of CMC documents are available. To remedy this situation, the TEI SIG on CMC provides customised TEI schemas with additional elements and attributes for encoding CMC. In particular, a <post> element to model the basic building block of CMC communication, along with several attributes such as @replyTo, and @indentLevel, was introduced by the TEI CMC SIG. The French and German projects that contributed to the TEI CMC SIG and in which CMC-specific TEI customisations were developed, were concerned with building CMC corpora of multiple genres (Beißwenger et al. 2012, Chanier et al. 2014, Lüngen et al. 2016).

Our proposal is based on the latest schema version provided by the TEI CMC SIG, called the CLARIN-D schema for CMC, (Beißwenger et al. 2016). In the following, we use the @replyTo and @indentLevel attributes as customised by the TEI CMC SIG for the <post> element, as well as grouped <link> elements from the regular TEI for the representation of reply relations in TEI CMC documents.

4.2 Annotation proposal based on TEI and TEI CMC SIG

Remember that a reply relation instance always occurs between a post and a previous post within one CMC interaction. We propose to encode technical reply relations using the attribute @replyTo at the <post> element as customised in the CLARIN-D TEI schema for CMC (Listing 3).

Listing 3: Attributes of a post from a blog comment thread.

```
<post synch="#t046" who="#u012_waschke" xml:id="p007" replyTo="#p004">
```

We propose to use the attribute @indentLevel at the <post> element as customised in the CLARIN-D TEI schema to represent all indentation structures in Wiki talk (Beißwenger et al. 2016), regardless of whether they are to be interpreted as reply relations or not (Listing 7).

Finally, we newly propose to encode and collect all *interpreted* reply relations (whether based on technical reply, indentation, or linguistic markers) in the TEI header of the CMC document as a set of <link> elements gathered within a <linkGrp>. Table 1 summarizes the proposed annotations.

Table 1: Summary of proposed annotations for reply relations.

Type of reply relation	Proposed annotation	Source
Technical reply	use attribute @replyTo at <post>	Chat2CLARIN, cf. Beißwenger et al. (2016)
Indentation	use attribute @indentLevel at <post>	DeRiK, cf. Beißwenger et al. (2012)
Interpretative	use <link> elements in <linkGrp>	TEI Guidelines, http://www.tei-c.org/release/doc/tei-p5-doc/de/html/ref-link.html

According to the TEI Guidelines, a <link> element quite generally "define[s] an association or hypertextual link among elements or passages".[17] A <link> implies a set of targets in its @target attribute, i.e. pointers to those elements in the text that are to be linked (always a pair of post IDs in our application) (Listing 4).

17 http://www.tei-c.org/release/doc/tei-p5-doc/de/html/ref-link.html (accessed 12 January 2019).

Listing 4: Interpreted reply relations in the TEI header.

```
<teiHeader>
[...]
      <linkGrp>
      <link target="#p1 #h1" type="initial"/>
      <link target="#p2 #p1" type="implied"/>
      <link target="#p3 #p1" type="addressing"/>
      </linkGrp>
[...]
</teiHeader>
```

We argue that the right place for the links is a link group in the TEI header of the CMC document because firstly, it is nice to have all of them collected in one place, so that they can be easily evaluated. Secondly, we can also *type* the abstract reply links using regular TEI means, i.e. the @type attribute at the <link> element, such as to capture information about the source or reason of an interpreted reply relation; according to our examples, we suggest the possible values "technical", "indentation", "addressing", "QA-relation", "quoting", and "implied" for the time being, while in the case that different markers signal the same relation (Case B in Section 3), @type may also contain a list of value strings. Thirdly, the encoding via <link> references offers the possibility to encode multiple reply relations originating from one post if desired (e.g. Grumt Suárez et al. 2016), thus it has the potential to go beyond the proper tree structure of threads. Lastly, <link> references can even be applied to represent reply relations that occur between other parts of the CMC documents than posts, such as headings or paragraphs, or groups formed of these.

We propose that the <linkGrp> could go in the <correspDesc> element of the file description of the TEI header, a section originally introduced to include information about the addressing, sending and receiving actions concerning and epistolary document. We are open to alternative suggestions of placing the <linkGrp>, only so far we found the <correspDesc> the most reasonable location based on its semantics. However, since <linkGrp> is currently not part of the content model of <correspDesc> in the TEI, we customised this for our examples.

Listing 5: Part of the TEI document body for Example 2 with interpreted reply relations in the TEI header.

```
<teiHeader>
[...]
      <correspDesc type="interpretedReplyRelations">
            <linkGrp>
               <link target="#p2 #p1" type="QA-relation"/>
               <link target="#p3 #p1" type="QA-relation"/>
            </linkGrp>
      </correspDesc>
[...]
</teiHeader>
<text>
[...]
<post mode="written" generation="human" synch="#t001" who="#A02" xml:id="p1"> Kann einer von euch vll einen. Guten Friseur empfehlen :)? </post>
<post mode="written" generation="human" synch="#t002" who="#A03" xml:id="p2"> Oliver Schmidt ist aber etwas teurer </post>
<post mode="written" generation="human" synch="#t002" who="#A04" xml:id="p3"> Ich hab leider auch keinen. Gehe immer nur zu einer Freundin zum schneiden aktuell </post>
[...]
</text>
```

Listing 5 shows an extract of the TEI document body of the in Example 2 displayed WhatsApp conversation. The interpreted reply relations between the three messages is presented in the link group within the TEI header. The source of the relations (how they were signalled) is indicated in the values in the @type attributes as “QA-relation”.

Listing 6: Part of the TEI document body for Figure 7 with interpreted reply relations in the TEI header.

```
<teiHeader>
[...]
<correspDesc type="interpretedReplyRelations">
      <linkGrp>
         <link target="#p2 #p1" type="technical implied"/>
         <link target="#p3 #p2" type="technical QA-relation"/>
      </linkGrp>
</correspDesc>
[...]
</teiHeader>
<text>
[...]
<post mode="written" type="tweet" generation="human" synch="#tweets.t001" xml:id="p1"
who="#u1" xml:lang="deu"> <time generation="system"> 15:15 </time> Brückenschlag zum
Emoji-Vortrag auf der <ref type="hashtag" target="...">#XBK2018</ref>: Die
Ausschmückungsfunktion von Emojis ist in den Whatsapp-Nachrichten von meiner Mutter
definiv perfektioniert...[...]"</post>
<post mode="written" generation="human" type="tweet" who="#u2" synch="#tweets.t002"
replyTo="#p1" xml:lang="de" xml:id="p2"> <time generation="system"> 18:15 </time> Wie
lange es wohl dauert bis <ref type="hashtag" target="...">#gifs</ref> diese Aufmerksamkeit
von LinguistInnen, Medien- und BildwissenachaftlerInnen bei <ref type="hashtag"
target="...">#cmc</ref> bekommen? <ref type="hashtag" target="...">#Justwondering</ref>
<ref type="twitter-account" target="...">@XBK_2018</ref> <ref type="hashtag"
target="...">#XBK2018</ref> Ich nehme schon auch wahr, dass die immer beliebter werden.
Und ihr?</post>
<post mode="written" generation="human" type="tweet" who="#u3" synch="#tweets.t003"
replyTo="#p2" xml:id="p3" xml:lang="deu"> <time generation="system"> 19:06 </time> Hm,
gibt's schon: u.a. <ref type="link" target="...">https://doi.org/10.1080/
08351813.2016.1164391</ref> und <ref type="link" target="...">https://doi.org/10.1177/
1470357216645481</ref> und <ref type="link" target="...">https://doi.org/10.1145/
2615569.2615697</ref> <ref type="hashtag" target="...">#gifs</ref></post>
[...]
</text>
```

Listing 6 presents a part of the TEI document body for the Twitter example (Figure 7). In addition to the technical reply relation that is initiated by replying to the respective tweet, we observed that the reply relations indicated through the use of a technical reply are additionally signalled by linguistic markers. Therefore, the values of @type contain more than one source.

Listing 7: Part of the TEI document body for Figure 8 with interpreted reply relations in the TEI header.

```
<teiHeader>
[...]
<correspDesc type="interpretedReplyRelations">
      <linkGrp>
         <link target="#p1 #h1" type="initial"/>
         <link target="#p2 #p1." type="implied"/>
         <link target="#p3 #p1" type="addressing"/>
      </linkGrp>
</correspDesc>
[...]
</teiHeader>
<text>
[...]
<div type="thread">
      <head xml:id="h1">Fehler bei Sprachgruppen</head>
      <post xml:id="p1" who="#u001" synch="#t001" indentLevel="0">
         <p>73,80 % sind deutscher Muttersprache!!!!!!!!!!!!!!!!!![...]</p></post>
      <post xml:id="p2" who="#u002" synch="#t002" indentLevel="0">
         <p>Nein, das stimmt schon so [...]</p></post>
      <post xml:id="p3" who="#u003" synch="#t003" indentLevel="1">
         <p><ref type="addressingTerm" target="#u001">Hi IP</ref>, [...]</p></post></div>
[...]
</text>
```

Listing 7 displays annotations of the Wiki talk example in Figure 8. The presented comments can be interpreted as examples of different types of reply relations. We said above that contrary to its indentation structure, post *p1* relates back to the thread heading *h1*. This is a common feature in Wiki talk pages as the users describe their requests efficiently in this way. The post *p1* is unindented ("indentLevel=0"), i.e. it can be interpreted as an initial post which relates to the overall topic stated in the thread heading (*h1*). Even though user 1 does not finish his post (*p1*) with a question, it becomes clear when looking at the content that the post *#2* is to be interpreted as a reply to *p1*.

5 Conclusion

In this paper, we have shown various types of reply relations that exist between post units in computer-mediated communication. We classified three types of re-

ply relations in CMC interactions according to how they are signalled: technical replies, indentations, and interpretative reply relations. We have outlined a combined representation of reply relations within the TEI framework. We use the @replyTo and @indentLevel attributes as customised by the TEI CMC SIG for the <post> element, as well as grouped <link> elements from the regular TEI in the TEI header. We claim that the presented annotation scheme would constitute a good base level for higher level annotations of interaction analysis such as dialogue acts (Ferschke et al. 2012), or discussion trees (Laniado et al. 2011). The described examples show the variety of interaction and response structures that can be found in CMC genres such as Twitter, blog, WhatsApp and Wiki talk. The users versatilely apply different addressing strategies in order to react to a given post. Besides a formally signalled reply relation, other linguistic cues can be found that support or override the existing formal reply. Distinguishing different levels of reply relations can thus be useful in many projects that have only relied on the formal reply relations so far.

References

Beißwenger, Michael. 2007. *Sprachhandlungskoordination in der Chat-Kommunikation*. Berlin. New York: de Gruyter (Reihe Linguistik – Impulse & Tendenzen 26).

Beißwenger, Michael, Maria Ermakova, Alexander Geyken, Lothar Lemnitzer & Angelika Storrer. 2012. DeRiK: A German Reference Corpus of Computer-Mediated Communication. *Proceedings of Digital Humanities 2012*. 259–263.

Beißwenger, Michael, Eric Ehrhardt, Axel Herold, Harald Lüngen & Angelika Storrer. 2016. (Best) Practices for Annotating and Representing CMC and Social Media Corpora in CLARIN-D. *Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities*. 7–11.

Chanier, Thiery, Celine Poudat, Benoit Sagot, Georges Antoniadis, Ciara Wigham, Linda Hriba, Julien Longhi & Djamé Seddah. 2014. The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres. In Michael Beißwenger, Nelleke Oostdijk, Angelika Storrer & Henk van den Heuvel (eds.), *Building and Annotating Corpora of Computer-Mediated Communication: Issues and Challenges at the Interface of Corpus and Computational Linguistics*. Special Issue, Journal of Language Technology and Computational Linguistics (JLCL 2/2014).

Ferschke, Oliver, Iryna Gurevych & Yevgen Chebotar. 2012. Behind the Article: Recognizing Dialog Acts in Wikipedia Talk Pages. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. 777–786.

Grumt Suárez, H., Natali Karlova-Bourbonus & Henning Lobin. 2016. A Discourse-structured Blog Corpus for German: Challenges of Compilation and Annotation. In Michael Beißwenger, Michael Wojatzki & Torsten Zesch (eds.), *NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication* (Bochumer Linguistische Beiträge 17), 1–5. Bochum: University of Bochum.

Ho-Dac, Lydia-Mai, Veronika Laippala, Céline Poudat & Ludovic Tanguy. 2016. French Wikipedia Talk Pages: Profiling and Conflict Detection. *Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities*. 34–38.

Holmer, Torsten. 2008. Discourse structure analysis of chat communication. *Language@Internet* 5. Article 9.

Laniado, David, Riccardo Tasso, Yana Volkovich & Andreas Kaltenbrunner. 2011. When the Wikipedians Talk: Network and Tree structure of Wikipedia Discussion Pages. *International AAAI Conference on Web and Social Media, Fifth International AAAI Conference on Weblogs and Social Media*. 177–184.

Lüngen, Harald, Michael Beißwenger, Eric Ehrhardt, Axel Herold & Angelika Storrer. 2016. Integrating corpora of computer-mediated communication in CLARIN-D: Results from the curation project ChatCorpus2CLARIN. In Stefanie Dipper, Friedrich Neubarth & Heike Zinsmeister (eds.), *Proceedings of the 13th Conference on Natural Language Processing* (KONVENS 2016). (= Bochumer Linguistische Arbeitsberichte (BLA) 16), 156–164.

Margaretha, Eliza, Harald Lüngen. 2014. Building Linguistic Corpora from Wikipedia Articles and Discussions. *Journal of Language Technology and Computational Linguistics*. 59–82.

Poudat, Céline, Jin Kun & Thierry Chanier. 2014. Wikiconflits, un corpus extrait de Wikipédia: principe et méthode d'élaboration. In Céline Poudat, Natalia Grabar, Jin Kun & Camile Paloque-Berges (eds.), *Corpus Wikiconflits, conflits dans le Wikipédia francophone*. Banque de corpus CoMeRe. Ortolang.fr: Nancy. [cmr-wikiconflits-tei-v4.1-manuel.pdf; http://hdl.handle.net/11403/comere/cmr-wikiconflits].

Schmidt, Thomas. 2011. A TEI-based approach to standardising spoken language transcription. Selected Papers from the 2008 and 2009 TEI Conferences. *Journal of the Text Encoding Initiative* 1 [https://journals.openedition.org/jtei/142].

Schröck, Jasmin, Harald Lüngen. 2015. Building and Annotating a Corpus of German-Language Newsgroups. In Michael Beißwenger & Torsten Zesch (eds.), *NLP4CMC 2015. 2nd Workshop on Natural Language Processing for Computer-Mediated Communication / Social Media. Proceedings of the Workshop*. 17–22.

TEI Consortium. 2019. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 3.5.0. [https://tei-c.org/guidelines/p5/].